Pareto depth for functional data

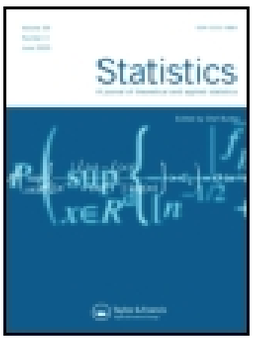
Helander, Sami; Van Bever, Germain; Rantala, Sakke; Ilmonen, Pauliina

Published in:
Statistics

DOI:
10.1080/02331888.2019.1700418

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record



Taylor & Francis
Taylor & Francis Group

Statistics
A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: https://www.tandfonline.com/loi/gsta20

Pareto depth for functional data

Sami Helander, Germain Van Bever, Sakke Rantala & Pauliina Ilmonen

To cite this article: Sami Helander, Germain Van Bever, Sakke Rantala & Pauliina Ilmonen (2019): Pareto depth for functional data, Statistics, DOI: 10.1080/02331888.2019.1700418

To link to this article: https://doi.org/10.1080/02331888.2019.1700418

STATISTICS
https://doi.org/10.1080/02331888.2019.1700418

Taylor & Francis
Taylor & Francis Group

Pareto depth for functional data

Sami Helander[a], Germain Van Bever[b], Sakke Rantala[c] and Pauliina Ilmonen[a]

[a]Department of Mathematics and Systems Analysis, Aalto University School of Science, Aalto, Finland; [b]Department of Mathematics & Namur Center for Complex Systems (naXys), Université de Namur, Namur, Belgium; [c]Kemijoki OY, Rovaniemi, Finland

ABSTRACT

This paper introduces a new concept of depth for functional data. It is based on a new multivariate Pareto depth applied after mapping the functional observations to a vector of statistics of interest. These quantities allow to incorporate the inherent features of the distribution, such as shape or roughness. In particular, in contrast to most existing functional depths, the method is not limited to centrality only. Properties of the depths are explored and the benefits of a flexible choice of features are illustrated on several examples. In particular, its excellent classification capacity is demonstrated on a real data example.

ARTICLE HISTORY
Received 2 May 2019
Accepted 29 November 2019

KEYWORDS
Functional data analysis; Pareto optimality; statistical depth

1. Introduction

With the increase in precision of measurements and of storage capacity, the last two decades have seen a tremendous jump in the dimensionality of data available. One of the common methodologies used when dealing with such high-dimensional observations is to assume that the observed units are random functions rather than random vectors. The pervasiveness of functional data in modern research – from stock prices to brain imaging, meteorology, or telecommunications – stemmed for a rigorous development of theories and methods for processing and analysing such data. As a result, many statistical methods such as (linear) regression, principal component analysis, canonical correlation, etc. have already been extended to functional settings (see, e.g. [1–3]).

The concept of statistical depth (see, e.g. [4–6]) was originally introduced as a way to palliate the absence of universal notion of quantiles in $\mathbb{R}^d$ and provide a centre-outward ordering from a depth-based multivariate median. In finite-dimensional setups, depth is a widely used, nonparametric, analytic tool. It does not only provide a measure of centrality but also reveals numerous features of the underlying distribution, such as asymmetry, spread or shape [7].

It is not surprising that a lot of attention was devoted to extending depth notions to functional setups, where modeling is known to be difficult and nonparametric approaches are common.

CONTACT Sami Helander sami.helander@aalto.fi Department of Mathematics and Systems Analysis, Aalto University School of Science, PO Box 11100, Aalto 00076, Finland

Most of the functional depths provided in the literature belong to one of two classes. The first approach typically integrates some centrality measure over the domain of the observations. This is the case, for example, for the integrated depth [8], the (modified) band depth [9], the (modified) half region depth [10], the integrated dual depth [11], the multivariate functional halfspace depth [12] or the general definition in [13]. The second class of definitions consists of notions that measure an expected distance from the function x to the distribution P. It includes the h-mode depth [14] and the functional version of spatial depth [15,16].

All the approaches above, however, are focused on examining the – pointwise – centrality of the functions as a measure of their (global) centrality in the distribution P. As a result, they are missing some features inherent to functional data such as shape, roughness or range. This is with the exception of the multivariate functional halfspace depth which also takes into account the derivative of the function. As it still proceeds with a pointwise integration, it misses global features. Note that some recent works aim at detecting – specific – types of (shape) outlyingness in functional data. They include, for example, [17–20].

The aim of this paper is to provide a depth notion that takes into account different characteristics of the functions to build a global functional depth definition. Due to the richness of functional data, we believe it is impossible to provide a turnkey, universal, notion that would fit any type of data. The approach adopted here is therefore to assume that mappings from the functional space to $\mathbb{R}^d$, that quantify some inherent features of the distribution, are provided beforehand. These mappings, called henceforth *statistics of interest* (SOI), will give the discriminating components within the distribution. Our construction then proceeds by computing a new multivariate depth, the *Pareto Depth*, on the vector of SOI.

Our approach is very much in line with classical FDA methods which use functional principal component analysis (FPCA) to project the distribution onto a finite dimensional subspace and proceed with multivariate methods. FPCA, however, assumes that the interesting features are obtained via a linear projection, which might not necessarily be the case in general. The approach taken here is, in that sense, more flexible and allows to put emphasis on shape, roughness, nonlinear or non-integrated characteristics of the functions.

In the functional context we described, the geometry of the random vector of SOI is irrelevant as each component bears its own meaning. The multivariate Pareto depth used is then purposely built on componentwise comparison of the chosen typicality measures. It is defined using Pareto level sets on the SOI. Intuitively, a functional observation will be deep (i.e., typical, rather than central) if its statistics of interest are close to their medians.

The paper is organized as follows: Section 2 introduces the general framework adopted in this paper. It describes the Hilbert space setting used throughout as well as a generic approach to depth functions and the notion of statistics of interest. Section 3 then introduces the (multivariate and functional) Pareto depths, while Section 4 gives refined versions of them. Their properties are studied in Section 5 while Section 6 illustrates their excellent behaviour in practice. Section 7 concludes with a short discussion on future prospects. Proofs are collected in the Appendix.

2. General framework

The functional framework adopted here assumes that the data points are random realizations in a Hilbert space $\mathcal{H}$. The choice of $\mathcal{H}$ is vast and depends on the type of data observed. However, the coined term 'functional' originally assumed $\mathcal{H}$ to be a set of functions on $\mathcal{V}$, a compact subset of $\mathbb{R}^d$, satisfying one or several regularity conditions and equipped with an appropriate inner product. One such space – by far the most studied and assumed in the literature – is

$$\mathcal{H} = L^2(\mathcal{V}, \mathbb{R}),$$

the set of real-valued square integrable functions on $\mathcal{V}$ with respect to some dominating measure ν, henceforth assumed to be the Lebesgue measure, with the inner product

$$\langle \cdot, \cdot \rangle : L^2(\mathcal{V}, \mathbb{R}) \times L^2(\mathcal{V}, \mathbb{R}) \to \mathbb{R} : (X, Y) \mapsto \langle X, Y \rangle = \int_{\mathcal{V}} X(t) Y(t) \nu(\mathrm{d}t).$$

More formally, we observe a *random function* X, that is, a measurable mapping $X : \Omega \to \mathcal{H}$ from the probability space $(\Omega, \mathcal{A}, P)$ to $(\mathcal{H}, \mathcal{B})$, where $\mathcal{B}$ is the σ-field generated by the open sets induced by the norm $\langle \cdot, \cdot \rangle$. For clarity, the terminology random function will be used throughout, regardless of the underlying Hilbert space $\mathcal{H}$, which could be non-functional.

While all the concepts introduced in this paper are valid in a general Hilbert space framework, the Sobolev space $W^{k,2}(I, \mathbb{R})$, for some appropriate k and I, a closed interval in $\mathbb{R}$, will be used for illustration purposes. Recall that $W^{k,2}(I, \mathbb{R})$ is the space of Lebesgue square integrable functions on I whose weak derivatives up to order k are also square integrable.

Other spaces considered in the literature include (i) restricting $\mathcal{H}$ by adding further smoothness conditions (continuous, C^k, etc.; see [21]), (ii) more general Hilbert spaces, such as multivariate functional spaces [22] and functional manifolds [23], or (iii) separable metric spaces (see, e.g. [24]). Most examples found in the literature are equipped with an inner product, though.

One important feature of functional data in practice is the fact that X, being infinite dimensional, cannot be fully observed. The first step of many FDA algorithms therefore consists in smoothing the discretely-measured sample. The reconstruction of the functional observations has been discussed profusely in the literature (see, e.g. [1,2]). While this is a crucial part in any method, we assume in the following expository sections to have a fully observed sample $X_1, \ldots, X_n$ of i.i.d. random functions in $\mathcal{H}$. Smoothing will be conducted in Section 6 together with the simulations and the real data example.

A *depth function* $D : \mathbb{R}^d \to \mathbb{R} : x \mapsto D(x, P)$ associates to each $x \in \mathbb{R}^d$ a measure of its centrality with respect to the multivariate distribution P. The more central x is in P, the higher is its depth value. Following the founding concept of halfspace depth introduced by [5],

$$D_H(x, P) = \inf_{u \in S^{d-1}} P[u'(X - x) \geq 0], \tag{1}$$

where S^{d-1} denotes the hypersphere in $\mathbb{R}^d$, numerous other depth functions were defined in the following decades (see, among others, [6,25]).

Several definitions extending the concept of depth to the functional setting have been introduced in the literature. They all aim at providing a function

$$D(\cdot, P) : \mathcal{H} \to \mathbb{R} : x \mapsto D(x, P)$$

measuring how *adequate* x is with respect to P. However, while several authors (see, e.g. [12]) pointed out the fact that centrality might not be the sole characteristic of interest in the functional setup, most of the definitions do not take into account such things as shape, roughness, etc. As a result, a central curve that differs, for example, only in its shape pattern from the rest of the data might be, counter-intuitively, considered deep.

Parallel to the halfspace depth construction – which aggregates projected outlyingness on each direction of the hypersphere – adopting a *projection approach* to functional depth might be of value. Cuevas et al. [14], for example, consider random projections of functional observations to define their depth. While this approach extends (1) to the functional setting, a random selection of elements in $\mathcal{H}$ to project upon does not guarantee an accurate selection of the important features of the data. Mosler and Polyakova [26] also consider projections by assigning to a function the minimal (multivariate) depth of its projections over a class Φ of $\mathbb{R}^d$-valued linear maps.

Alternatively, many methodologies in FDA replace X with the (random) vector $(\langle X, f_i\rangle_{\mathcal{H}}, i = 1, \ldots, d)$, where f_i is the ith eigenfunction of the covariance operator

$$\Sigma_P : \mathcal{H} \times \mathcal{H} \to \mathbb{R} : (x, y) \mapsto \int_{\mathcal{H}} \big(\langle x, z\rangle_{\mathcal{H}} \langle y, z\rangle_{\mathcal{H}}\big)\, \mathrm{d}P(z).$$

In the context of depth, one could then apply any multivariate depth function to the resulting random vector. However, this again amounts to assuming that the interesting characteristics are obtained (i) via linear projections and (ii) are linked to the second-order moment structure of P. These two assumptions are strong hypotheses in the functional case.

Centrality (i.e., location) and inner products with fixed elements in $\mathcal{H}$ (albeit data dependent) are not the only characteristics a depth function should measure in functional settings. Other features (spread, peakedness, etc.) might indeed be of interest in assessing the *typicality* of an element in $\mathcal{H}$. Moreover, the set of relevant features depend on the data at hand and the ability to consider various criteria of typicality is a great asset that helps shedding light on the behaviour of the underlying process from which the functional data has been observed.

Hence, it is natural to assume that, for a given measure P, statistics of interest (SOI) have been defined and selected. Let

$$T : \mathcal{H} \to \mathbb{R}^d : x \mapsto (T_1(x), \ldots, T_d(x)) \tag{2}$$

be the mapping that associates each element of $\mathcal{H}$ to its vector of SOI. Numerous SOI can be considered and the choice, naturally, depends on the underlying space $\mathcal{H}$. For $x(t) \in W^{k,2}(I, \mathbb{R})$, possible measures include

(i) $T(x) = \int_I x(t)\, \mathrm{d}t$ (centrality);
(ii) $T(x) = \int_I x'(t)\, \mathrm{d}t$ (shape);

(iii) $T(x) = \max_I x(t) - \min_I x(t)$ (spread);
(iv) $T(x) = \|x\|_{W^{k,2}}$ (roughness/peakedness).

This very flexible approach not only allows different foci to be put on P according to what one judges typical in the data but also enables experts to provide ad hoc SOI. This will be the case, for example, in a water level dataset studied in Section 6. In general, we suggest using a vector of SOI capturing at least location, shape and roughness.

Once random functions have been mapped to $\mathbb{R}^d$, a multivariate depth can be applied to the obtained vector of SOI. The next section introduces the notion of Pareto depth. While the whole section is applicable to a general distribution P, the reader should keep in mind that the primary target of this approach is to be applied to the vector of SOI. The functional depth obtained by applying the Pareto depth to the vector of SOI will henceforth be called *Functional Pareto depth*.

3. Pareto depth

Similar to convex hull peeling depth [27,28], where depth is defined by recursively peeling the boundary of the convex hull of the data, (multivariate) Pareto depth assigns depth from the centre outwards, by recursively peeling the Pareto optimal observations.

For ease of presentation, we first restrict to the empirical setting, where the (multivariate) dataset $\mathcal{T} = \{T_1, \ldots, T_n\} \subset \mathbb{R}^d$ is observed. Let $T_i = (T_{i1}, \ldots, T_{id})'$ and $P_{\mathcal{T}}$ denote the empirical distribution on $\mathcal{T}$. Note that, in the functional framework, T_{ik} represents the score of the ith random function on the kth SOI. Let $s = (s_1, \ldots, s_d)' \in \mathbb{R}^d$ and let

$$f_k(s) = |s_k - \mathrm{med}_k(P_{\mathcal{T}})|, \quad k = 1, \ldots, d, \tag{3}$$

where, for e_k the kth canonical vector of $\mathbb{R}^d$, $\mathrm{med}_k(P)$ denotes the median of the (univariate) distribution $e_k'P$ (without loss of generality, we will assume throughout that the median is uniquely defined by, if necessary, averaging over the median set). The Pareto regions and the Pareto rank of $s \in \mathbb{R}^d$ are based on the notion of Pareto optimality and defined in the following way.

Definition 3.1: The point s is Pareto optimal in the set A if there is no element a in A such that $f_k(a) \leq f_k(s)$ for all $k \in \{1, \ldots, d\}$ with strict inequality for at least one f_k.

Definition 3.2: The Pareto regions $\mathcal{T}_\ell \subset \mathcal{T}_0 = \mathcal{T}, \ell = 1, \ldots, (L_{\mathcal{T}} + 1)$ are defined recursively via

$$\mathcal{T}_\ell = \mathcal{T}_{\ell-1} \setminus \mathcal{J}_\ell,$$

where $\mathcal{J}_\ell \subseteq \mathcal{T}_{\ell-1}$ is the set of observations that are Pareto optimal in $\mathcal{T}_{\ell-1}$. $L_{\mathcal{T}}$ is fixed such that $\mathcal{T}_{L_{\mathcal{T}}} \neq \emptyset$ and $\mathcal{J}_{L_{\mathcal{T}}+1} = \mathcal{T}_{L_{\mathcal{T}}}$. The set $\mathcal{J}_\ell, \ell = 1, \ldots, (L_{\mathcal{T}} + 1)$ is called the Pareto level set of order ℓ.

The Pareto regions are constructed by sequentially peeling the Pareto level sets. In particular, the Pareto region $\mathcal{T}_{(L_{\mathcal{T}}+1)}$ in the above definition is always the empty set $\varnothing$. Conventionally, s is always Pareto optimal with respect to $\varnothing$.

Definition 3.3: The Pareto rank of $s \in \mathbb{R}^d$ in $\mathcal{T} = \{T_1, \ldots, T_n\}$ is

$$Rank_{\mathcal{T}}(s) = \min \left\{ r \middle| s \text{ is Pareto optimal in } \mathcal{T}_r \right\}.$$

That is, the rank of s is r if s is Pareto optimal in the subset of observations obtained after peeling r times the Pareto optimal set of the dataset.

We are now ready to define the Pareto depth.

Definition 3.4: The Pareto depth of $s \in \mathbb{R}^d$ in $\mathcal{T} = \{T_1, \ldots, T_n\}$ is

$$PD_{(n)}(s, \mathcal{T}) = 1 - \frac{\text{Rank}_{\mathcal{T}}(s)}{(L_{\mathcal{T}} + 1)}.$$

As $\text{Rank}_{\mathcal{T}}(s) \in \{0, \ldots, L_{\mathcal{T}} + 1\}$, it is straightforward to see that $0 \leq PD_{(n)}(s, \mathcal{T}) \leq 1$. Moreover, $PD_{(n)}(s, \mathcal{T}) = 1$ if s is Pareto optimal in $\mathcal{T}$ and $PD_{(n)}(s, \mathcal{T}) = 0$ if, for all $k = 1, \ldots, d$, $f_k(s) > \max_i |T_{ik} - \text{med}(T_{ik})|$. Note that the definition of sample Pareto depth $PD_{(n)}(s, \mathcal{T})$ involves $L_{\mathcal{T}} + 1$, the number of level sets in $\mathcal{T}$. The peeling process must therefore be carried out in its entirety in order to find the Pareto Depth at a point s.

In order to define a population version of $PD(\cdot, \mathcal{T})$, one would need to extend the notion of Pareto ranks defined above to a continuous distribution P. Intuitively, one could proceed with sequentially peeling from $\mathbb{R}^d$ the Pareto optimal sets with respect to the componentwise distances to the marginal medians of P. For a continuous P, however, such peeling cannot be achieved as only the first level Pareto optimal set is non-empty.

Indeed, let P be a continuous distribution with strictly positive density at its componentwise median and let $X \sim P$. Then, the Pareto level set of order 1 contains only the componentwise median m_P of P. However, due to continuity, the infimum

$$\inf_{\mathbb{R}^d \setminus \{m_P\}} f_k(s) = 0 \quad \forall k = 1, \ldots, d,$$

is never achieved. Thus, the Pareto level set of order 2 is the empty set. It is therefore impossible to Pareto peel $\mathbb{R}^d$ with respect to a continuous distribution.

One way to provide a meaningful extension is to consider a *random sample* of i.i.d. observations $\mathcal{T} = \{T_1, \ldots, T_m\}$ with $T_i \sim P$.

Definition 3.5: The Pareto depth of $s \in \mathbb{R}^d$ of order m in the distribution P,

$$PD_m(s, P) = E[1 - R(s)],$$

where $R(s)$ is the random variable with value $Rank_{\mathcal{T}}(s)/(L_{\mathcal{T}} + 1)$, where $\mathcal{T} = \{T_1, \ldots, T_m\}$ is a set of m i.i.d. random vectors from P.

Rather than providing multiple rankings, each associated with a univariate SOI, Pareto depth allows to provide a global typicality ordering based on the joint vector of SOI. The definition above is close in spirit to the convex hull probability depth from [29]. The depth is indeed also turned into a population version taking into account the discreteness that is necessary in taking a peeling approach. The main differences in the approach taken here – on top of looking at Pareto optimal regions rather than peeling the convex hull – are

(i) the fact that the depth is defined on the ranks rather than the probability content of the peeled regions and, more importantly, (ii) that the use of SOI allows to suppress the need to study the geometry of the multivariate distribution. Contrary to the convex hull peeling approach, the Pareto depth is ad hoc to the 'projections' used and allows for better interpretation and statistical properties, which is seen in later sections. Moreover, such construction makes the depth easy to evaluate, even in high dimensions. Pareto depth is therefore computationally feasible in settings where other geometric approches (such as halfspace depth) are not. This Pareto approach also ensures that the depth always ranges from zero to one.

The sample version of $PD_m(\cdot, P)$, obtained by plugging in $P = P_n$, approximates the original sample depth $PD_{(n)}(s, \mathcal{T})$ when m is large, as proved in the following result.

Theorem 3.1: *Let P_n be the empirical distribution on a fixed finite set $\mathcal{T} = \{T_1, \ldots, T_n\}$. Then, for any $s \in \mathbb{R}^d$,*

$$\lim_{m\to\infty} \sup_{s\in\mathbb{R}^d} \left| PD_m(s, P_n) - PD_{(n)}(s, \mathcal{T}) \right| = 0.$$

All proofs can be found in the Appendix. The proof of Theorem 3.1 shows that, for m big enough, the sampling of m points from P_n covers $\mathcal{T}$ and concludes with the equality of the two depth functions. Consequently, $PD_m(\cdot, P)$ can be seen as a continuous version of $PD_{(n)}(\cdot, \mathcal{T})$.

4. Pareto depth refined

Depth is about measuring centrality of observations. Each component can deviate from the median to some extent. However, should deviation in only one of the marginals be enough to give a point a low depth value?

An inherent feature of the Pareto depth introduced above is that any observation with only one marginal close to the median will be given a large depth value, despite the fact that its other marginals might have very untypical values. This would, in particular, include observations in the deepest regions that only exhibit a central pattern in a small subset of its components.

To address this, a modification of the Pareto depth function is introduced. It ensures that the deepest observations are actually most central across large enough subsets of the SOI, that is of the marginals of T. The definition includes a parameter $\lambda \in \mathbb{N}$ which allows to tune the number of marginals in which s should be central in order to be associated with a high depth. Intuitively, the modified λ-Pareto depth of s is the minimal Pareto depth of s over subsets of marginals of size $d - \lambda$. The sample version is provided in the definition below. The population version follows along the same lines as Definition 3.5.

Definition 4.1: Let $\lambda \in \mathbb{N}$. The λ-Pareto depth of $s \in \mathbb{R}^d$ in the set $\mathcal{T} = \{T_1, \ldots, T_n\}$,

$$PD^{(\lambda)}_{(n)}(s, \mathcal{T}) = \min_{\{i_1,\ldots,i_\lambda\}=\mathcal{I}\subset\{1,\ldots,d\}} PD_{(n)}(s^{-\mathcal{I}}, T^{-\mathcal{I}}),$$

where $s^{-\mathcal{I}}$ denotes the $(d - \lambda)$-vector s from which components $s_{i_1}, \ldots, s_{i_\lambda}$ were removed and $T^{-\mathcal{I}} = \{T_i^{-\mathcal{I}}, i = 1, \ldots, n\}$.

The choice of λ depends on d and affects the resulting Pareto depth ordering of the dataset. It holds that $PD^{(0)}_{(n)}(s, \mathcal{T}) = PD_{(n)}(s, \mathcal{T})$. Intuitively, $(1 + \lambda)$ is the size of the subset of marginals with respect to which s must be central to reach a high depth. Further discussion on the choice of the parameter value λ as well as its significance in practice is provided in Section 6.

We close this section with a formal definition of functional Pareto depth.

Definition 4.2: The functional λ-Pareto depth of order m for the SOI function T is

$$FPD^{(\lambda)}_m(., P) : \mathcal{H} \to \mathbb{R} : x \mapsto FPD^{(\lambda)}_m(x, P) = PD^{(\lambda)}_m(T(x), P_T),$$

where P_T denotes the distribution of $T(X)$, for $X \sim P$.

5. Properties

The Pareto depth function $PD_m(\cdot, P)$ can be shown to have many desirable properties. The following theorem states its consistency.

Theorem 5.1: *Let P be a distribution on $\mathbb{R}^d$. Let $T_1, \ldots, T_n$ be i.i.d. observations from P. Let P_n be their empirical distribution. For any $s \in \mathbb{R}^d$ and $m \in \mathbb{N}$, as $n \to \infty$,*

$$PD_m(s, P_n) \xrightarrow{P} PD_m(s, P).$$

Note that $PD_m(s, P_n)$ is a random variable that considers m observations taken with replacement from the random set $\{T_1, \ldots, T_n\}$. Theorem 5.1 then simply states the weak convergence of $PD_m(s, P_n)$ to its mean.

In the multivariate setting, [4] defined a statistical depth function as a function $D(\cdot, P)$ that fulfills the following four properties:

(P1) Affine invariance: $D(As + b, P_{AX+b}) = D(s, P_X)$ for any non-singular $d \times d$ matrix A and $b \in \mathbb{R}^d$, where X has distribution P_X and P_{AX+b} denotes the distribution of $AX + b$.
(P2) Maximality at centre: If P is symmetric about $\theta \in \mathbb{R}^d$, then $D(\theta, P) = \max_s D(s, P)$.
(P3) Monotonicity along rays from the deepest point: If there is a deepest point s_0, then $D(s_t, P)$ is monotonically decreasing along any ray $s_t = s_0 + tu$, for all $t \in \mathbb{R}$, $u \in S^{d-1}$.
(P4) Vanishing at infinity: $D(s, P) \to 0$ as $\|s\| \to \infty$.

Building on componentwise optimization of distance functions, the Pareto depth is not affine-invariant (it is, trivially, translation-invariant, though). However, due to the nature of the construction in the functional setting, the geometry of the random vector of SOI is not relevant. Only componentwise centrality of SOI is of importance. Thus, lack of invariance is not a concern as the marginals in that case have their own meaning. Properties (P2) and (P3), on the other hand, do have their importance in the context of Pareto depth and are shown to hold in Theorem 5.2. Property (P4), however, does not hold. Indeed, if $\|s\| \to \infty$ is such that $s_k = \text{med}_k(P)$ for some $k \in \{1, \ldots, d\}$, then $PD(s, P) = 1$, for all s. Nonetheless, the weaker property (P4')

(P4') Vanishing at infinity: $D(s,P) \to 0$ as $\min\{s_1,\ldots,s_d\} \to \infty$

is also shown to hold in the next theorem.

Theorem 5.2: *The Pareto Depth function $PD_m(s,P)$ satisfies* (P2) (*for halfspace symmetric distributions*), (*P*3) *and* (*P*4′).

It is also interesting to study the properties that the functional Pareto depth exhibits. Axiomatic approaches for functional depth have been developed by Nieto-Reyes and Battey [24] and Gijbels and Nagy [30]. Providing a general study, including the continuity (or upper semi-continuity) of functional Pareto depth, however, would go beyond the expository nature of this manuscript. The properties indeed depend on the choice of statistics of interest (function) S, m, λ, and $\mathcal{H}$. As an illustration, Theorem 5.3 adopts the same notations as in [30] and explores the axiomatic properties of $FPD_m^{(\lambda)}(.,P)$ for $\lambda = 0$ and $\mathcal{H} = L^2(\mathcal{V},\mathbb{R})$.

Theorem 5.3: *Fix $\mathcal{H} = L^2(\mathcal{V},\mathbb{R})$ and $m \in \mathbb{N}_0$. Let T be an SOI function and P be a distribution on $\mathcal{H}$. The functional Pareto depth $FPD_m^{(0)}$ satisfy the following properties:*

(P-0) *Non-degeneracy: Under the assumption that P_T is absolutely continuous with respect to the Lebesgue measure,*

$$\inf_{x\in\mathcal{H}} FPD_m^{(0)}(x,P) < \sup_{x\in\mathcal{H}} FPD_m^{(0)}(x,P).$$

(P-1) *Invariance: Assume that T preserves the Pareto ordering under a class of functions $\mathcal{C} = \{f : \mathcal{H} \to \mathcal{H}\}$, that is, $Rank_{\mathcal{T}}(T(x)) = Rank_{f(\mathcal{T})}(T(f(x)))$ for any $x \in \mathcal{H}, f \in \mathcal{C}$ and $\mathcal{T} = \{x_1,\ldots,x_m\} \subset \mathcal{H}$. Then, for all $f \in \mathcal{C}$,*

$$FPD_m^{(0)}(f(x),P_{f(X)}) = FPD_m^{(0)}(x,P).$$

(P-2) *Maximality at centre: Assume that T is odd. Then, for any P with a unique centre of symmetry $\theta \in \mathcal{H}$ (for some notion of functional symmetry),*

$$FPD_m^{(0)}(\theta,P) = \max_{x\in\mathcal{H}} FPD_m^{(0)}(x,P).$$

(P-3) *Decreasing from the deepest point: If T is linear, for any P such that $FPD_m^{(0)}(z,P) = \sup_{x\in\mathcal{H}} FPD_m^{(0)}(x,P)$, for any $\alpha \in [0,1]$, for any $x \in \mathcal{H}$,*

$$FPD_m^{(0)}(x,P) \le FPD_m^{(0)}(x+\alpha(z-x),P).$$

(P-4) *Vanishing at infinity: Under the assumption that P_T is absolutely continuous with respect to the Lebesgue measure, $FPD_m^{(0)}(x,P) \to 0$ as $\|x\| \to \infty$.*

Note that the assumptions on T in Theorem 5.3 are satisfied for all or any combination of the four statistics of interest described in Section 2 (with the exception of linearity for $T(x) = \|x\|_{W^{k,2}}$). In particular, property P-1 holds for the class $\mathcal{C}$ of all transformations that are monotonely increasing in each marginal. The assumptions of Theorem 5.3 will also be satisfied in the simulations section below.

6. Simulations

In this section, we consider four simulated and one real dataset examples. All the examples highlight the need for a flexible functional depth able to consider characteristics beyond location.

The functional depths considered in this section cover recent proposals belonging to both categories described in the introduction. On top of the functional Pareto depth (FPD) described above, the following depths – known to outperform their competitors in most cases – will be used:

(i) the multivariate functional halfspace depth (MFHD, [12]), applied to the functional observations and their derivatives; and
(ii) the kernelized functional spatial depth (KFSD, [31]), with a gaussian kernel and the automatic bandwidth selection provided by the authors.

Simulated examples The first two simulated examples consider a single centrally placed shape outlier, where as the other two consider group contamination. Examples 6.1 and 6.2 were generated from a smooth gaussian process following the method suggested in [32], page 14–16, with covariance operator

$$\mathrm{Cov}(x(t_s), x(t_r)) = \kappa(t_s, t_r) = \exp\left(-\frac{1}{2}(|t_s - t_r|/l)^2\right),$$

where the characteristic length-scale l specifies how fast the values of $x(t)$ are allowed to vary.

In each replication, a set of $n = 25$ observations was generated. Each observation was evaluated on the interval $[-5, 5]$ over a total of 50 equidistant measurement points. The sample paths of the observations were conditioned to run through the 10 training values

$$\{(-4, -2), (-3.5, -1), (-3, 0), (-2, 0.5), (-1, 1), (0, 2), (1, 0), (2, -1), (3, 0), (4, 1)\},$$

with allowed standard deviation $\sigma = 0.25$ from the training points and with $l = 0.6$.

For both datasets, an outlier was then generated and added to the observations, yielding datasets of size $N = 26$. Two types of outliers were considered.

Example 6.1 (Centrally located smooth outlier): The outlier was generated using the same method as above with a denser set of centrally placed training values, no allowed variance, and an increased length-scale.

Example 6.2 (Centrally located rough outlier): The outlier was generated using the same method as in Example 6.1 except that the 50 measured values were then perturbed by adding the vector $(0, 0.2, 0, -0.2, 0, \ldots, 0.2)$.

Examples 6.3 and 6.4 consider two sets of curves from different base models with outliers added from contaminating distributions. These models were previously analysed by López-Pintado and Romo [9] and Sun and Genton [33].

Example 6.3 (Peaked contamination): The non-contaminated curves follow the base model $X_i(t) = 4t + e_i(t)$, $i = 1, \ldots, n_X$, where $e_i(t)$ is a stochastic gaussian process with

zero mean and covariance function $\gamma(t_s, t_r) = (\frac{1}{2})(\frac{1}{2})^{5|t_r - t_s|}$. The outliers were generated from the model $Y_j(t) = 4t + e_j(t) + \mathbb{1}_{\{t \in [L_j, L_j + l]\}} \sigma_j M$, $j = 1, \ldots, n_Y$, where $e_j(t)$ is a zero mean stochastic gaussian process as in the base model, M and l are constants determining the height and width of the peaks respectively, σ_j is a sequence of random variables taking values 1 and -1 with probability 1/2, L_j is a random number from a uniform distribution on $[0, 1 - l]$, and $\mathbb{1}$ is the indicator function.

Example 6.4 (Noise contamination): The base curves follow the model $X_i(t) = 4t + e_{xi}(t)$, $i = 1, \ldots, n_X$, with $e_{xi}(t)$ a zero mean gaussian stochastic process with covariance function $\gamma_1(t_s, t_r) = \exp(-|t_r - t_s|^2)$. The outliers were generated from the model $Y_j(t) = 4t + e_{yj}(t), j = 1, \ldots, n_Y$, with $e_{yj}(t)$ a zero mean gaussian stochastic process with covariance function $\gamma_2(t_s, t_r) = \exp(-|t_r - t_s|^{0.2})$.

In both Examples 6.3 and 6.4, $n_X = 45$ observations were generated from the base model with $n_Y = 5$ outliers added from the contaminating model, yielding datasets of size $N = 50$.

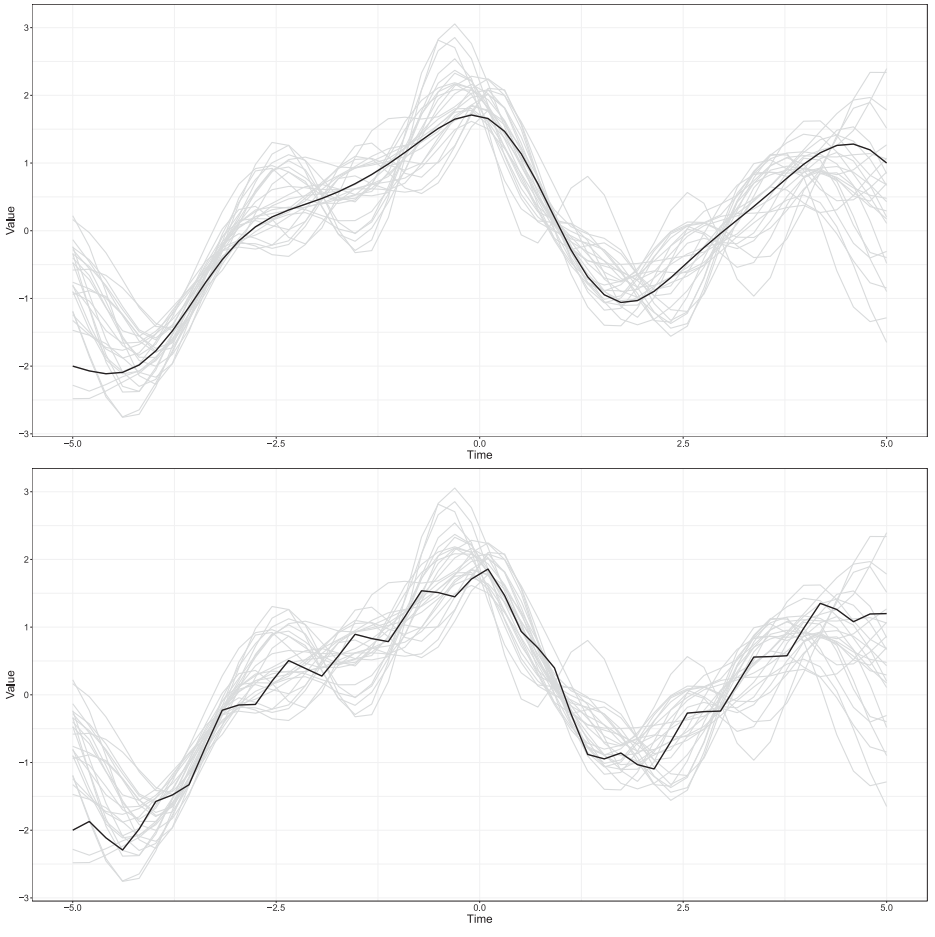


Figure 1. $N = 26$ observations from the gaussian process described above with a smooth (top) and a rough (bottom) outlier.

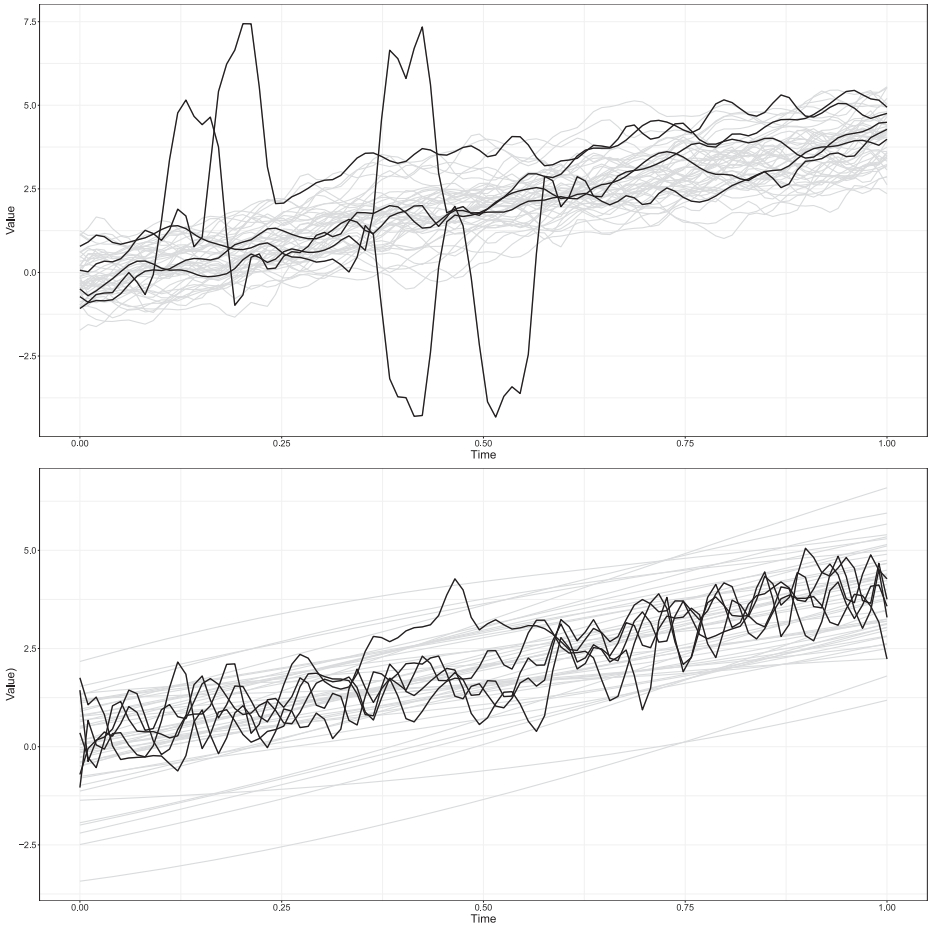


Figure 2. $N = 50$ observations from Examples 6.3 (top) and 6.4 (bottom). The outliers are highlighted in black.

In each example, a set of functional data was obtained by smoothing the simulated vectors of measurement values, using a B-spline basis of order 4 with 50 equidistantly placed knots. Note that, in particular, the outlying observations in all four setups are indeed continuously differentiable up to degree 4. All examples in this section were generated and smoothed using the R package 'fda'.

Figure 1 presents simulated datasets from Examples 6.1 and 6.2, while Figure 2 presents simulated datasets from Examples 6.3 and 6.4. In both figures the outlying observations are highlighted in black.

Functional depths of the observations in each dataset were then calculated. Functional Pareto depth was based on the following statistics of interest:

(1) Location: $T_{i1} = \int \sum_{j=1}^{N} \text{sign}(x_i(t) - x_j(t))(x_i(t) - x_j(t))^2 \, \mathrm{d}t$
(2) Averaged roughness: $T_{i2} = \int |x_i''(t)| \mathrm{d}t$.
(3) Number of zero derivatives: $T_{i3} = \#\{t : x_i'(t) = 0, \quad \nexists \epsilon > 0 : x_i'(t^*) = 0 \quad \forall t^* \in [t - \epsilon, t]\}$

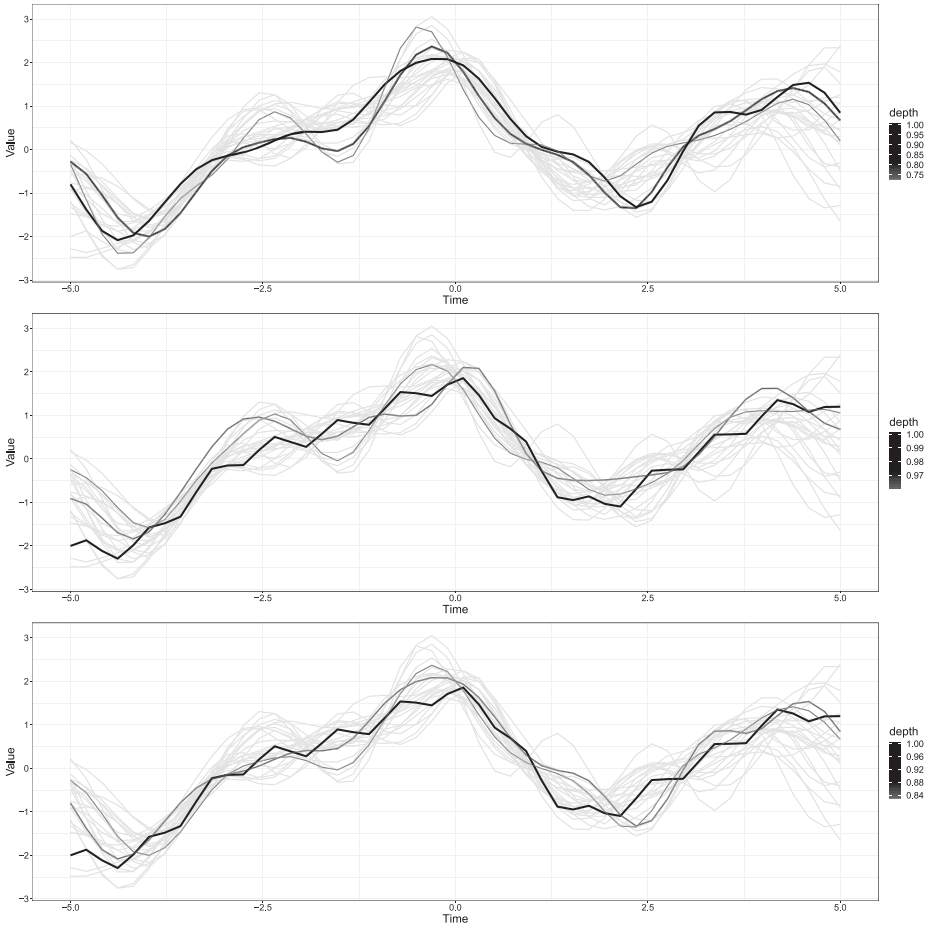


Figure 3. The dataset used in the bottom panel of Figure 1. The three deepest curves (Pareto depth: top; MFHD: middle; KFSD: bottom) are highlighted.

The location measure T_{i1} is a *signed* centrality measure, integrated pointwise over time, and taking positive or negative values. Centrally-located observations are such that $|T_{i1}| \approx 0$. For functional data, typicality in the shape of an observation is of equally great importance as location. Due to the variance functions used in the examples, the simulated observations were allowed to vary quite rapidly. T_{i2} and T_{i3} were used to measure how frequently an observation oscillates, as well as the amplitude of that oscillation. These two indicators combined provide a good proxy of the typicality of the shape of an observation.

In all four simulated examples, the same set of statistics of interest was used to highlight the relative ease of choosing SOI that meaningfully capture the key features of a distribution.

In all simulated examples the parameter $\lambda = 1$ was used and the Pareto depth values $PD^{(1)}_{(n)}(T_i, \mathcal{T})$, for $\mathcal{T} = \{T_1, \ldots, T_N\}$ and $T_i = (T_{i1}, T_{i2}, T_{i3})$, were computed. The choice of λ is such that, in order to be associated with high depth, an observation is required to be typical in at least two of the used SOI.

Table 1. Median depths and ranks of the outlying observation over 100 replications of Examples 6.1 and 6.2.

	Example 6.1			Example 6.2		
	Pareto	MFHD	KFSD	Pareto	MFHD	KFSD
Depth	0	0.92	0.74	0	0.75	0.65
Rank	2.5	25	22	2.5	23	20

Table 2. Median depths and ranks of the outlying observation with the highest depth, over 100 replications of Examples 6.3 and 6.4.

	Example 6.3			Example 6.4		
	Pareto	MFHD	KFSD	Pareto	MFHD	KFSD
Depth	0.16	0.57	0.10	0.05	0.08	0.45
Rank	12	31	5	7.5	10	25

Figure 3 highlights the 3 deepest curves of the dataset displayed in the bottom panel of Figure 1 for the three depth functions considered. The highlighted observations are visually similar between MFHD and KFSD. Notably, both MFHD and KFSD rank the outlying curve among the deepest observations whereas Pareto depth does not. For Pareto depth, the highlighted observations are not only location-typical but also shape-typical.

For all four examples, 100 independent datasets were generated. For Examples 6.1 and 6.2, the depth and rank of the outlying curve was computed. For Examples 6.3 and 6.4 the depths and ranks of all 5 outliers were computed for each replication, from which the the outlier with the highest depth was recorded. Tables 1 and 2 report the median values for each of the three depths considered. Note that depth values were rescaled to the unit interval to allow for better comparison.

In the first two examples, MFHD and KFSD consistently give a high depth value to the outlying curve and rank it among the deepest observations. This is the case in Example 6.1, which, admittedly is a difficult outlying detection problem since the entire dataset, not just the central outlier, is rather smooth in nature. However, this is also the case in Example 6.2, where the roughness of the outlying curve differs drastically from the rest of the data. On the other hand, Pareto depth consistently flags the outlying curve as an atypical observation. Indeed, in both examples, the median Pareto depth value is 0.

In the latter two examples, MFHD and KFSD exhibit very different performances, each ranking the outlying observations low in one of the examples, but failing in the other. KFSD picks out peaked outliers in Example 6.3 while MFDH does relatively well with the rough outliers of Example 6.4. Pareto depth performs well in both examples (and best in Example 6.4), flagging most of the outliers as atypical, even with a generic choice of SOI. Note that choosing the SOI in a more case-specific way (such as range for Example 6.3) would yield even better results. However, to achieve fair comparison between simulation settings, the same generic vector of SOI were adopted.

To conclude this simulation section, Table 3 provides the average computational times (over 50 replications, using R 3.6.1 on an Intel i5-4690K 3.5 GHz processor) of Pareto depth and halfspace depth for $N = 1000$ standard gaussian observations in dimension $d = 2, 3, 4, 5$. Following the R 'depth' package, halfspace depth was approximated for $d \geq 3$.

Table 3. Comparison of computational times (in seconds), taken as an average over 50 replications, of halfspace depth (D_H) and multivariate Pareto Depth ($PD^{(\lambda)}$) over a simulated d-variate dataset with $N = 1000$ random standard gaussian observations.

d	D_H	$PD^{(0)}$	$PD^{(1)}$	$PD^{(2)}$	$PD^{(3)}$
2	0.16	0.13	0.18	–	–
3	7.72	0.28	0.4	0.34	–
4	8.99	0.53	1.04	0.69	0.58
5	10.63	0.85	2.4	2.17	1.16

Table 3 clearly illustrates the benefits from using a Pareto projection approach over a geometric depth. Note that Pareto depth remains computable in high dimensions when halfspace depth is notoriously impossible to compute in dimensions larger than 5.

Real data example: Kemijoki The Kemijoki dataset[1] depicts the water reservoir surface levels of three hydro power plants on the Kemijoki river that are not mutually directly connected. The measurements are taken as a difference from the maximum level in meters. For simplicity, these differences are henceforth referred to as 'levels'. The data consists of observations from 484 different days drawn from multiple years, with 144 measurements for each day. The data is presented in Figure 4 with randomly chosen observations highlighted in solid black line to help perceive the overlapping observations.

Reservoirs A and C behave similarly having a rather stable surface level. Their daily means (average level during a given day) are close to each other and have an average value of -0.23 and -0.24, respectively. The most notable differences between the two reservoirs are the slightly tighter grouping of the majority of the data in Reservoir C as well as its outlying observations behaving clearly differently with drastically fluctuating surface levels. Additionally, Reservoir B is clearly distinguishable from the others in both location and shape. Its observations are located around the daily mean of -0.15 and exhibit much less overlap with drastically lower standard deviations of the daily means (0.035 compared to 0.072 and 0.079 for Reservoirs A and C, respectively).

The daily measurements were smoothed using a B-spline basis of order 4 with knots placed at each measurement point. The resulting functional observations $x_i(t)$, $i = 1, \ldots, n = 484$ interpolate the data almost exactly.

As mentioned in previous sections, the contextual knowledge provided by an expert of the field can prove invaluable in determining the relevant features of interest. In this particular case, the production process of the plant is used as a mean to adjust the power grid to the fluctuations in demand during a day. The following statistics of interest were suggested by an expert from Kemijoki OY:

(1) Daily mean: $T_{i1} = \int x_i(t)\,\mathrm{d}t$.
(2) Range: $T_{i2} = \max_t x_i(t) - \min_t x_i(t)$.
(3) Averaged absolute fluctuations: $T_{i3} = \int |x_i'(t)|\,\mathrm{d}t$.
(4) Averaged roughness: $T_{i4} = \int |x_i''(t)|\,\mathrm{d}t$.
(5) Number of daily oscillations: $T_{i5} = \#\{t : x_i'(t) = 0,\ \nexists \epsilon > 0 : x_i'(t^*) = 0\ \forall t^* \in [t - \epsilon, t]\}$.

The rationale behind the choice of these SOI is the following. The interest lies in the full power production behaviour during a given day. While the mean level is a good proxy for

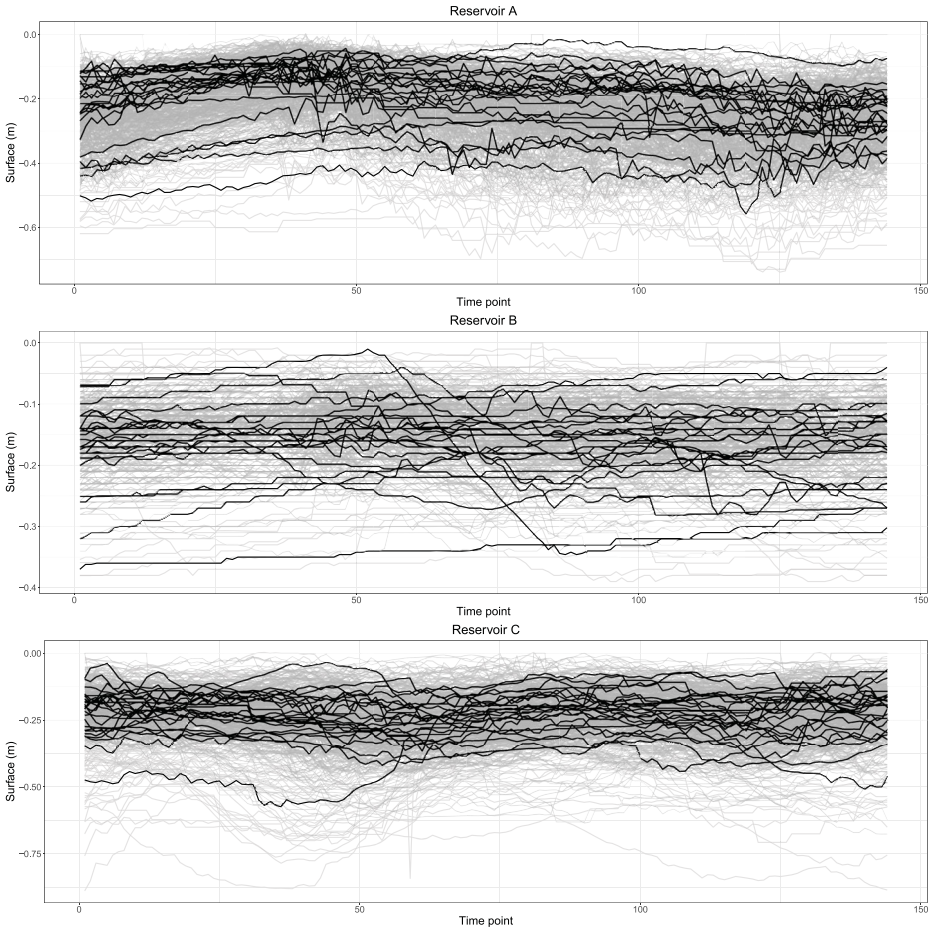


Figure 4. Hydro power plant reservoir levels with randomly chosen observations highlighted in black.

the typicality of the day globally, it doesn't analyse the behaviour of the production during the day. Indeed, the surface level is expected to fluctuate amply as water is discharged from the reservoir. Days with very stable level are deemed untypical, even if they have a very typical mean level. Therefore, the second and third SOI in combination reveal features of the power production during a day, by measuring the flow of water through the reservoirs. Furthermore, another point of interest lies in the peakedness versus continuity of the daily power production. When power is being produced continuously, water is constantly discharged from the reservoir resulting in smooth changes in the reservoir levels. However, intermittent power production results in peaks in the surface level data, the number and amplitude of which the fourth and fifth SOI measure.

The choice of λ reflects the analyst's perception of how many SOI values should be simultaneously typical in order for the curve to be considered deep. While λ is impactful on the depth values and their distribution, it does not, in this case, drastically change the ordering they provide. Restricting to $\lambda > 0$ offers the required flexibility to the depth function as it now associates observations performing well in only one SOI to lower depth values. In the

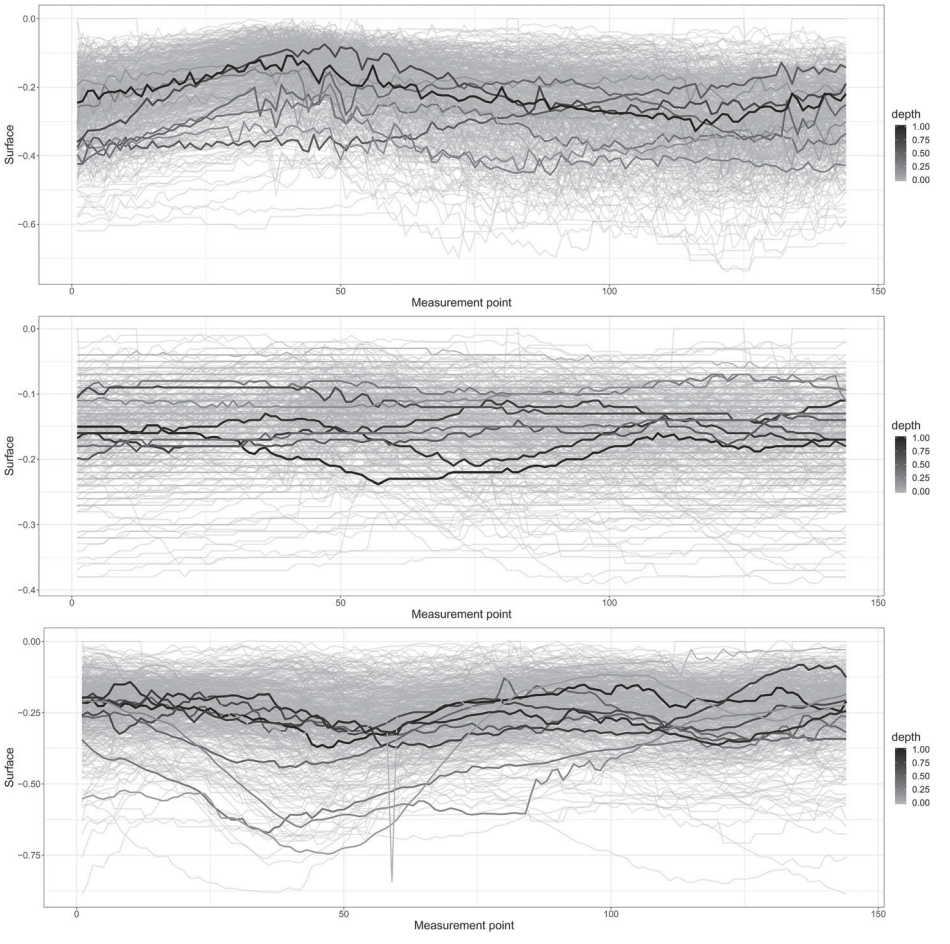


Figure 5. Pareto Depth values for selected observations of Kemijoki water level data.

case of Kemijoki dataset, the values $\lambda = 1$ and $\lambda = 4$ lead to heavily skewed distributions of the depth values, making them unappealing choices. On the other hand $\lambda = 2$ and $\lambda = 3$ lead to depth distributions with little or no skewness. As some SOI are naturally paired, since they aim at measuring different aspects of the same phenomenon, the choice $\lambda = 3$ is more natural. Indeed, in that case, an observation actually has to perform well in 4 components in order to be deemed of high depth. Note that, in practice, λ is the only parameter that needs to be selected. The parameter m appearing in Section 3 is of importance only for theoretical purposes. As a general guideline, we suggest to use different values of λ to assess the sensitivity of the depth values to subsets of SOI.

The Pareto depth values $PD^{(3)}_{(484)}(T_i, \mathcal{T})$, for $\mathcal{T} = \{T_1, \ldots, T_{484}\}$ and $T_i = (T_{i1}, \ldots, T_{i5})$ are illustrated in Figure 5. For each reservoir, the entire dataset has been drawn in the background and 10 observations have been highlighted and greyscale-coloured according to their depth value. The highlighted observations were chosen uniformly from the ordered dataset, where the ordering was based on the corresponding depth values. The depth values have been rescaled to the unit interval for ease of comparison.

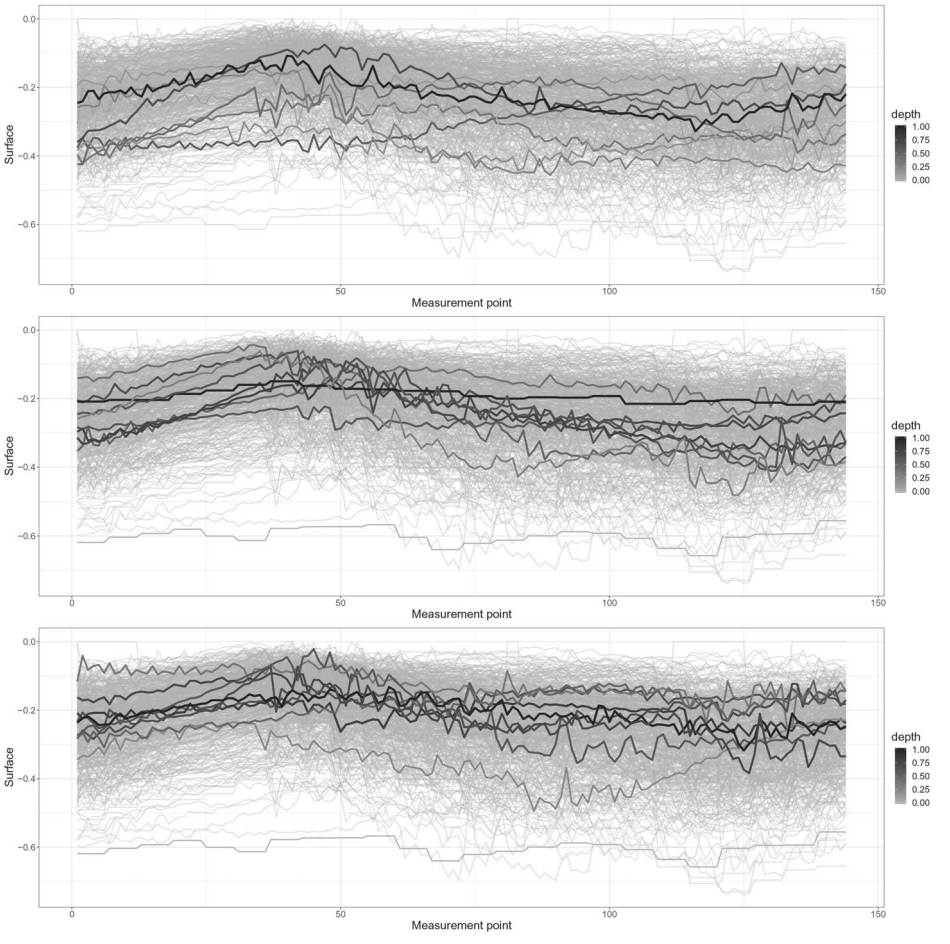


Figure 6. Depth values (Pareto: top; MFHD: middle; KFSD: bottom) for selected observations of Reservoir A of Kemijoki water level data.

A comparison of the three functional depths (Pareto depth, MFHD and KFSD) for Reservoir A is presented in Figure 6. The curves are displayed as in Figure 5.

In the figures, the impact of the SOI on the Pareto depth value is immediately apparent, especially when it comes to the shape of an observation. Centrality, while clearly an important factor, now plays a visibly smaller role compared to MFHD or KFSD. Instead, the qualities deemed important in the context of hydro power production can be identified in the observations with high depth, and the lack of some of these qualities becomes apparent as the depth of an observation decreases.

To determine to which extent the functional depths are able to capture the essential features in different reservoirs, maximum depth classification [34] was performed between the reservoirs.

A leave-one-out classification scheme was conducted for each pair of reservoirs. Sequentially, each observation was taken out of the pooled sample and classified in the reservoir

Table 4. Leave-one-out misclassification rates for each reservoir pair based on max-depth classication with Pareto Depth ($PD^{(\lambda)}$), MFHD, KFSD and based on *k*NN.

	$PD^{(1)}$	$PD^{(2)}$	$PD^{(3)}$	$PD^{(4)}$	MFHD	KFSD	*k*NN
AvB	0.036	0.029	0.023	0.032	0.229	0.249	0.156
BvC	0.057	0.038	0.036	0.048	0.214	0.207	0.173
CvA	0.212	0.155	0.127	0.157	0.262	0.243	0.174

with respect to which it had the highest depth value. In case of ties, the reservoir was chosen at random.

Max depth classification was conducted for each possible pair of reservoirs and for each depth function (Pareto Depth, MFHD, KFSD). In addition, a *k nearest neighbours* (KNN) classification [35] was performed. The classification was based on L^2 distances. For each reservoir pair, k was chosen by leave-one-out crossvalidation over the set $k \in \{3, 5, 7, \ldots, 483\}$. This lead to the choices of $k_{AB} = 5$, $k_{BC} = 3$ and $k_{CA} = 17$.

The leave-one-out misclassification rates are presented in Table 4. MFHD and KFSD perform equally well across each classification problem, with *k*NN reaching a slightly better performance. Pareto depth clearly outperforms the other methods for $\lambda > 1$, and still performs favourably with $\lambda = 1$.

7. Final comments

This paper provides a new multivariate depth as well as an original functional depth based on the former. The functional version obtained by peeling sequentially the Pareto levels of the vectors of statistics of interest. Most importantly, this new approach is not only examining location but incorporates information about shape, roughness, etc. to assess the typicality of a curve.

The choice of statistics of interest is guided by prior knowledge on the observations. Naturally, there are many applications in which there is no such a priori information on where the typicality of the observations lies. While there is of course no perfect choice of SOI, we believe that choosing a set of functions measuring the three aspects of centrality, shape and roughness is an excellent candidate. Most importantly, depth functions, in general, are exploratory tools designed to shed light on the underlying structure of the data and the proposition in this paper fall into this framework.

Future theoretical challenges, outside the scope of this expository paper, include (i) quantifying the effect of the choice of λ in $PD_m^{(\lambda)}(s, P)$, (ii) exploring further the properties of the functional Pareto depth with respect to the choice of SOI and (iii) study the geometry of the Pareto depth regions and understand how they characterize the underlying distribution.

Note

1. The data is shared by Kemijoki Oy to the scientific community for academic research purposes by the original request of Department of Mathematics and Systems Analysis at Aalto University, Finland. Any other use of the data is not allowed. Due to possible competitive advantage reasons, any distinguishing information of the data, including the dates and specific reservoirs, have been removed. The data is not publicly available, but can be redistributed for research purposes on request.

Acknowledgments

The authors are grateful to two anonymous referees for their careful reading and insightful comments that led to substantial improvements of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Ramsay J, Silverman B. Functional data analysis. New York: Springer-Verlag; 2005.
[2] Ferraty F, Vieu P. Nonparametric functional data analysis: theory and practice. New York: Springer-Verlag; 2006.
[3] Horváth L, Kokoszka P. Inference for functional data with applications. New York: Springer-Verlag; 2012.
[4] Zuo Y, Serfling R. General notions of statistical depth function. Ann Statist. 2000;28(2): 461–482.
[5] Tukey JW. Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2. Canad. Math. Congress, Montreal, Que.; 1975. p. 523–531
[6] Liu RY. On a notion of data depth based on random simplices. Ann Statist. 1990;18(1):405–414.
[7] Liu RY, Parelius JM, Singh K. Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion). Ann Statist. 1999;27(3):783–858.
[8] Fraiman R, Muniz G. Trimmed means for functional data. Test. 2001;10(2):419–440.
[9] López-Pintado S, Romo J. On the concept of depth for functional data. J Amer Statist Assoc. 2009;104(486):718–734.
[10] López-Pintado S, Romo J. A half-region depth for functional data. Comput Statist Data Anal. 2011;55(4):1679–1695.
[11] Cuevas A, Fraiman R. On depth measures and dual statisticsA methodology for dealing with general data. J Multivariate Anal. 2009;100:753–766.
[12] Claeskens G, Hubert M, Slaets L, et al. Multivariate functional halfspace depth. J Amer Statist Assoc. 2014;109:411–423.
[13] Hlubinka D, Gijbels I, Omelka M, et al. Integrated depth for functional data: statistical properties and consistency. ESAIM: Probab Statist. 2016;20(1):95–130.
[14] Cuevas A, Febrero M, Fraiman R. Robust estimation and classification for functional data via projection-based depth notions. Comput Statist. 2007;22:481–496.
[15] Chakraborty A, Chaudhuri P. On data depth in infinite dimensional spaces. Ann Inst Statist Math. 2014;66:303–324.
[16] Chakraborty A, Chaudhuri P. The spatial distribution in infinite dimensional spaces and related quantiles and depths. Ann Statist. 2014;42:1203–1231.
[17] Nagy S, Gijbels I, Hlubinka D. Depth-based recognition of shape outlying functions. J Comp Graph Statist. 2017;26(4):883–893.
[18] Dai W, Genton M. Directional outlyingness for multivariate functional data. Comput Statist Data Anal. 2019;131:50–65.
[19] López-Pintado S, Ying S, Lin J, et al. Simplicial band depth for multivariate functional data. Adv Data Anal Classi. 2014;8(3):321–338.
[20] Rousseeuw PJ, Raymaekers J, Hubert M. A measure of directional outlyingness with applications to image data and video. J Comp Graph Statist. 2018;27(2):345–359.
[21] Hsing T, Eubank R. Theoretical foundations of functional data analysis, with an introduction to linear operators. Chichester: Wiley; 2015.
[22] Chiou J, Müller HG. Linear manifold modelling of multivariate functional data. J R Stat Soc Ser B Stat Methodol. 2014;76(3):605–626.

[23] Chen D, Müller HG. Nonlinear manifold representations for functional data. Ann Statist. 2012;40(1):1–29.
[24] Nieto-Reyes A, Battey H. A topologically valid definition of depth for functional data. Statist Sci. 2016;31(1):61–79.
[25] Oja H. Descriptive statistics for multivariate distributions. Statist Probab Lett. 1983;1(6): 327–332.
[26] Mosler K, Polyakova Y. General notions of depth for functional data. Statistics and Econometrics, Universität zu Köln; 2016. ArXiv:1208.1981v2
[27] Barnett V. The ordering of multivariate data (with discussion). J R Stat Soc Ser A General. 1976;139:318–352.
[28] Donoho D, Gasko M. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. Ann Statist. 1992;20:1803–1827.
[29] Porzio G, Ragozini G. On some properties of the convex hull probability depth. Cassino: Department of Economics, University of Cassino; 2010.
[30] Gijbels I, Nagy S. On a general notion of depth for functional data. Statist Sci. 2017;32(4):630–639.
[31] Sguera C, Galeano P, Lillo R. Functional outlier detection by a local depth with application to NOx levels. Stoch Environ Res Risk Assess. 2016;30(4):1115–1130.
[32] Rasmussen C, Williams C. Gaussian processes for machine learning. Cambridge, MA: MIT Press; 2006.
[33] Sun Y, Genton M. Functional boxplots. J Comput Graph Statist. 2011;20(2):316–334.
[34] Ghosh AK, Chaudhuri P. On maximum depth and related classifiers. Scand J Statist. 2005;32(2):327–350.
[35] Biau G, Bunea F, Wegkamp M. Functional classification in Hilbert spaces. IEEE Trans Inf Theory. 2005;51:2163–2172.

Appendix. Proofs

This section details the proofs of Theorems 3.1, 5.1 and 5.2.

Proof of Theorem 3.1**.:** By definition of PD_m, we have that

$$PD_m(s, P_n) = \frac{1}{n^m} \sum_{i_1,\ldots,i_m=1}^{n} \left(1 - \frac{Rank_{\mathcal{T}_{(i_1,\ldots,i_m)}}(s)}{(L_{\mathcal{T}_{(i_1,\ldots,i_m)}} + 1)}\right),$$

where $\mathcal{T}_{(i_1,\ldots,i_m)}$, for the ordered m-tuple $(i_1, \ldots, i_m)$, is the set $\{T_{i_1}, \ldots, T_{i_m}\}$ with possible repetitions. Let

$$\mathcal{M} = \left\{(i_1, \ldots, i_m) | i_j \in \{1, \ldots, n\}\right\}$$

and, for $m \geq n$, let

$$A = \left\{(i_1, \ldots, i_m) \middle| \mathcal{T}_{(i_1,\ldots,i_m)} = \mathcal{T}\right\} \subset \mathcal{M} \quad \text{and} \quad B = \mathcal{M} \setminus A.$$

It now holds that $Rank_{\mathcal{T}_{(i_1,\ldots,i_m)}}(s) = Rank_{\mathcal{T}}(s)$ if $(i_1, \ldots, i_m) \in A$. The relative cardinality of B with respect to n^m, the total number of possible samples, is given by

$$\frac{1}{n^m} \sum_{k=1}^{n-1} \binom{n}{k} k^m, \tag{A1}$$

which converges to 0 as $m \to \infty$.

Fix $\epsilon > 0$. Let M be such that (A1)$< \epsilon/2$ if $m \geq M$. Let $T_{\mathcal{I}} = T_{(i_1,\ldots,i_m)}$ for $\mathcal{I} = (i_1, \ldots, i_m)$. It now follows that, for $m \geq M$ and for any $s \in \mathbb{R}^d$,

$$\begin{aligned}
\left|PD_m(s, P_n) - PD_{(n)}(s, \mathcal{T})\right| &= \left| \frac{1}{n^m} \sum_{\mathcal{I} \in A} \left(1 - \frac{Rank_{\mathcal{T}_{\mathcal{I}}}(s)}{(L_{\mathcal{T}_{\mathcal{I}}} + 1)}\right) \right. \\
&\quad \left. + \quad \frac{1}{n^m} \sum_{\mathcal{I} \in B} \left(1 - \frac{Rank_{\mathcal{T}_{\mathcal{I}}}(s)}{(L_{\mathcal{T}_{\mathcal{I}}} + 1)}\right) - PD_{(n)}(s, \mathcal{T}) \right| \\
&= \left| \frac{1}{n^m} \sum_{\mathcal{I} \in A} \left(1 - \frac{Rank_{\mathcal{T}}(s)}{(L_{\mathcal{T}} + 1)}\right) \right. \\
&\quad \left. + \frac{1}{n^m} \sum_{\mathcal{I} \in B} \left(1 - \frac{Rank_{\mathcal{T}_{\mathcal{I}}}(s)}{(L_{\mathcal{T}_{\mathcal{I}}} + 1)}\right) - \frac{1}{n^m} \sum_{\mathcal{I} \in \mathcal{M}} \left(1 - \frac{Rank_{\mathcal{T}}(s)}{(L_{\mathcal{T}} + 1)}\right) \right| \\
&= \left| \frac{1}{n^m} \sum_{\mathcal{I} \in B} \left(\left(1 - \frac{Rank_{\mathcal{T}_{\mathcal{I}}}(s)}{(L_{\mathcal{T}_{\mathcal{I}}} + 1)}\right) - \left(1 - \frac{Rank_{\mathcal{T}}(s)}{(L_{\mathcal{T}} + 1)}\right) \right) \right| \\
&< 2\frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

As M does not depend on s, we have that, for $m > M$,

$$\sup_{s \in \mathbb{R}^d} \left|PD_m(s, P_n) - PD_{(n)}(s, \mathcal{T})\right| < \epsilon.$$

This completes the proof. ■

Proof of Theorem 5.1***:*** Let $T_1, \ldots, T_n$ be i.i.d observations from P. Let

$$\mathcal{M} = \left\{(i_1, \ldots, i_m) | i_j \in \{1, \ldots, n\}\right\}.$$

For $\mathcal{I} = (i_1, \ldots, i_m) \in \mathcal{M}$, let $T_{\mathcal{I}}$ denote $\{T_{i_1}, \ldots, T_{i_m}\}$, with possible repetitions. The random variable $PD_m(s, P_n)$ writes as

$$PD_m(s, P_n) = \frac{1}{n^m} \sum_{\mathcal{I} \in \mathcal{M}} X_{\mathcal{I}}, \text{ with } X_{\mathcal{I}} = \left(1 - \frac{Rank_{\mathcal{T}_{\mathcal{I}}}(s)}{(L_{\mathcal{T}_{\mathcal{I}}} + 1)}\right).$$

It now follows immediately that $E(PD_m(s, P_n)) = PD_m(s, P)$. Furthermore,

$$Var(PD_m(s, P_n)) = \frac{1}{n^{2m}} \sum_{\mathcal{I} \in \mathcal{M}} \sum_{\mathcal{J} \in \mathcal{M}} Cov(X_{\mathcal{I}}, X_{\mathcal{J}}).$$

Note that $X_{\mathcal{I}}$ is independent from $X_{\mathcal{J}}$, for $\mathcal{I}, \mathcal{J} \in \mathcal{M}$ if none of their components are equal. For a fixed $\mathcal{I} \in \mathcal{M}$, when non-zero, $Cov(X_{\mathcal{I}}, X_{\mathcal{J}})$ can be bounded by $Var(X_{\mathcal{I}})$. The number of such non-zero quantities is

$$(n^m - (n - \mathcal{I}^\sharp)^m),$$

where $\mathcal{I}^\sharp$ is the number of components of $\mathcal{I}$ distinct from those in $\mathcal{J}$. This quantity can be bounded from above by $(n^m - (n-1)^m)$. Hence,

$$\begin{aligned}
Var(PD_m(s, P_n)) &\leq \frac{1}{n^{2m}} \sum_{\mathcal{I} \in \mathcal{M}} (n^m - (n-1)^m) Var(X_{\mathcal{I}}) \\
&= \frac{n^m - (n-1)^m}{n^m} Var(X_{\mathcal{I}}),
\end{aligned}$$

which converges to 0 as $n \to \infty$. This completes the proof. ■

Proof of Theorem 5.2**:** (P2): The property (P2) holds as, under symmetry, the componentwise median is the centre of halfspace symmetry, hence objective functions in (3) are all zero.

(P3): Along any ray from the componentwise median, the objective functions $f_k(s)$ in (3) are jointly non-decreasing in s. In that case, irrespective of the random dataset, (P3) holds.

(P4′): Let $\mathcal{T} = \{T_1, \ldots, T_m\}$ be a random sample of d-variate i.i.d. observations from P and let $T \sim P$. Let A_R denote the cube $\{(s_1, \ldots, s_d) | \, |s_k - med_k(P)| < R \; \forall k = 1, \ldots, d\}$.

For $\epsilon > 0$, let R_ϵ be such that $P(T \in A_{R_\epsilon}) \geq \sqrt[m]{1-\epsilon}$. For $T_i \sim P$, let I_i be the indicator that $T_i \in A_{R_\epsilon}$. It holds that

$$\begin{aligned} PD_m(s, P) &= E[1 - R(s)] \\ &= E\left[(1 - R(s))\left(\prod_{i=1}^{m} I_i\right)\right] + E\left[(1 - R(s))\left(1 - \prod_{i=1}^{m} I_i\right)\right]. \end{aligned}$$

For any $s \in \mathbb{R}^d$, the expectation $E[(1 - R(s))(1 - \prod_{i=1}^m I_i)]$ is bounded from above by

$$E\left[\left(1 - \prod_{i=1}^{m} I_i\right)\right] \leq 1 - P(T \in A_{R_\epsilon})^m = 1 - (\sqrt[m]{1-\epsilon})^m = \epsilon.$$

Assume that $\min(s_1, \ldots, s_d) > R_\epsilon + \max_k(|med_k(P)|)$. If $(\prod_{i=1}^m I_i) = 1$, it now follows that, for any $k \in \{1, \ldots, d\}$ and for any $j \in \{1, \ldots, m\}$

$$|s_k - med_k(P)| > \min(s_1, \ldots, s_d) - \max_k(|med_k(P)|) > R_\epsilon > |T_{jk} - med_k(P)|.$$

Thus, s is not Pareto optimal with respect to any $T_j \in \mathcal{T}$. Hence $Rank_{\mathcal{T}} = L_{\mathcal{T}} + 1$ and consequently $R(s) = 1$. Therefore, $(1 - R(s)) = 0$ if $(\prod_{i=1}^m I_i) = 1$.

Hence, if $\min(s_1, \ldots, s_d) > R_\epsilon + \max_k(|med_k(P)|)$, then $E[(1 - R(s))(\prod_{i=1}^m I_i)] = 0$ and $PD_m(s, P) < \epsilon$. This concludes the proof. ■

Proof of Theorem 5.3**:** The proof follows from Theorem 5.2 and the assumptions made on the statistics of interest. Note, for P-0, that non-degeneracy holds trivially for $PD_m^{(0)}$. Property P-1 follows from the invariance of depth under transformations preserving the Pareto ranks. P-3 follows from the fact that linearity preserves rays. ■